DOCUMENT RESUME

ED 319 590 SE 051 398

AUTHOR Shermis, Mark D.; St. George, Ross
TITLE Item Bias in Mathematics Achievement: The Progressive Achievement Tests for Mathematics.
PUB DATE Apr 90
NOTE 29p.; Paper presented at the Annual Meeting of the National Council for Educational Measurement (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Education; *Elementary School Mathematics; Elementary Secondary Education; Foreign Countries; Item Analysis; *Item Bias; *Item Response Theory; *Mathematics Achievement; *Mathematics Tests; Test Bias; Test Interpretation
IDENTIFIERS *New Zealand

ABSTRACT

This study examined differential item performance in the Progressive Achievement Tests for Mathematics (PAT Math) by both Pakeha (reference--majority) and Maori (focal--minority) students in New Zealand using both Item Response Theory (IRT) and non-IRT techniques. The PAT Math, a 50-item standardized test used in part to guide within class and school placement, was suspected of containing problems which might be biased in favor of the reference group. The sample consisted of the 726 Pakeha and 63 Maori students who had taken the test in 1985. After testing for unidimensionality, the data were calibrated using Logist V. Items were classified by patterns of Item Characteristic Curve (ICC) which were generated using the three-parameter logistic model. Six patterns were discovered, indicating different response patterns for the two groups. Results were compared against the Mantel-Haenszel chi-square test and the z-square test. Recommendations are made concerning the appropriate use of ICCs when the focal group size is too small to obtain stable IRT parameter estimates. The ICCs of all 50 items are appended. (Author/YP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED319590

Item Bias in Mathematics Achievement: The Progressive Achievement Tests for Mathematics

by

Mark D. Shermis

Department of Educational Psychology

The University of Texas at Austin

and

Ross St. George

Department of Psychology

Massey University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Mark D. Shermis

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

SE051398

Paper presented at the Annual Meetings of the National Council for Educational Measurement, Boston, MA, April 16-20, 1990.

Inquiries should be addressed to Mark D. Shermis, Department of Educational Psychology, EDB 504, The University of Texas at Austin, Austin, TX 78712-1296, (Voice) 512-471-4155/(Fax) 512-471-4607/(BITNET) EPFG745@UTXVM.

BEST COPY AVAILABLE

ABSTRACT

This study examined differential item performance in the *Progressive Achievement Tests for Mathematics* (*PAT Math*) by both Pakeha (reference--majority) and Maori (focal--minority) students in New Zealand using both IRT and non-IRT techniques. The *PAT Math (Form A)*, a 50 item standardized test used in part to guide within class and school placement, was suspected of containing problems which might be based in favor of the reference group. The sample consisted of the 726 Pakeha and 63 Maori students who had taken the test in 1985. After testing for unidimensionality, the data were calibrated using Logist V. Items were classified by patterns of ICC's which were generated using the three-parameter logistic model. Six patterns were discovered, indicating different response patterns for the two groups. Results were compared against the Mantel-Haenszel chi-square test and the z^2 test and recommendations were made concerning the appropriate use of ICCs when the focal group size is too small to obtain stable IRT parameter estimates.

Item Bias in Mathematics Achievement: The Progressive Achievement Tests for Mathematics

Introduction

The *Progressive Achievement Test (PAT)* series were developed and published in New Zealand between 1969 and 1978 by the New Zealand Council for Educational Research (NZCER). They are standardized norm and content-referenced educational achievement measures covering Reading Comprehension and Reading Vocabulary, Listening Comprehension, Mathematics and Study Skills. They are all paper-and-pencil, multiple-choice, multi-level tests in English based on the New Zealand school curriculum covering the student age range of 8 to 14 years.

Since their inception these tests have seen extensive use in New Zealand schools for instructional planning and placement purposes. They have also been primary quantitative data sources in numerous educational research projects. Because New Zealand is a multi-ethnic and multicultural society, the question of educational test suitability and test bias has been raised in relation to tests in the *PAT* series. The principle concern has been in relation to the poorer test performance of New Zealand Maori and Pacific Island Polynesians who constitute a minority of approximately 15 percent of the population.

These concerns and the outcomes of a series of studies into ethnic group differences in the *PAT*s have been reported by Beck and St. George (1983), and Chapman, St. George, and Ibell (1985). In response to allegations of test bias, studies of ethnic group differences in PAT item performance have been undertaken. These have been based on classical item analysis approaches (see St. George, 1987; St. George and Chapman, 1987). However, the recognized limitations of such approaches in equating the latent ability of the groups on the trait or characteristic in question suggests the application of between group item analysis procedures based on Item Response Theory (IRT).

This paper reports the outcomes from the first application of an item characteristic curve analysis of Maori and European (Pakeha) students on the *PAT Mathematics* test (Reid & Hughes, 1974).

Perspectives/Theoretical Framework

IRT has been used as a common tool for the study of item bias. Its major advantage stems from the invariance of measurement parameters which is assumed to hold for different subgroups. Deviations from this assumption are typically viewed as item bias (Muthen, Kao, & Burstein, 1988).

While the focus of recent item bias work has been on the explanation of bias (Scheuneman, 1987), the outcomes of such procedures in mathematics have been ambiguous. For example, Mellenbergh (1988) examined sets of modifications to a 60-item arithmetic test administered to Dutch (N=169) and Turkish/Moroccan (N=93) students at the end of the sixth grade. Initial item parameters were calculated based on 2500 Dutch and 451 Turkish/Moroccan students. Items were manipulated to reflect the changes in: (1) the plausibility of the response alternatives (10 items), (2) item placement (8 items), (3) wording difficulty (8 items), and (4) item complexity (12 items). The results indicated statistically significant changes in levels of bias for only three of the thirty-eight modified items. Other authors (e.g., Muthen, Kao, & Burstein, 1988) have chosen to test differences in item performance with a view towards controlling for the instructional background or 'opportunity-to-learn' of the student.

Data Sources

The *PAT Math (Form A)* is a 50 item mathematics achievement test which covers ten subdomains. The test has proven to be generally reliable (parallel forms range: r =.87-.88; KR-20 range: r =.89-.93) and correlates highly with other mathematics achievement tests (e.g., r =.82 with the *ITBS Modern Math Supplement*). The sample consists of 726 European-background and 63 Maori students at the Form 2 class level

(age range, 11.0 to 13.9 years-old). All students had completed *Form A, Part 6* of the *PAT Mathematics* tests. A full sample description and methodological description is available in St. George and Chapman (1987).

Method

The PAT (Mathematics-Form A) data (k=50) were subjected to an IRT analysis which consisted of an examination of the item characteristic curves (ICC) for both the Maori and Pakeha groups. The ICC for the three-parameter logistic model is represented in the following mathematical form:

$$P \equiv P(\theta) = c + \frac{1 - c}{1 + e^{-1.7a(\theta - b)}} \qquad [1]$$

where

$P(\theta)$ is the probability that an individual at ability θ will get the item correct,

a is the item's discrimination,

b is the item's difficulty, and

c is the pseudo-guessing parameter[1].

The use of the three-parameter logistic model was problematic in that 18 of the items were comprised of 4 alternatives while 32 of the items consisted of 5 alternatives. Instead of eliminating the four-alternative items from the analysis, it was decided to treat all items as if they had five choices from which students could choose.

IRT analysis assumes certain properties of the data. First it is predicated on the notion that the test matter is unidimensional. While *PAT (Mathematics)* purportedly measures mathematics achievement in several broadly defined areas, it

---

[1]Some authors refer to the c parameter as the *pseudo*-guessing parameter (cf. Hambleton & Swaminathan, 1985) because even though the probability of getting an item correct can be calculated *a priori* (e.g., five response alternatives = .20 probability of a correct response via random guessing), empirical results seldom match their theoretical predictions.

still may be thought of as tapping one overall dimension (Reid & Hughes, 1974). In order to test for unidimensionality of the 50 item test, the response data were factor analyzed using a principal components analysis. Input for the factor analysis consisted of a tetrachoric correlation matrix calculated by the SYSTAT 5.0™ statistical analysis program (Wilkinson, 1989). Factors with eigenvalues greater than one along with their variance contributions are presented in Table 1.

Table 1. Eigenvalues and proportion of total variance (Eigenvalues > 1.0) from the factor analysis of the Progressive Achievement Test (Mathematics--Form A, Part 6).

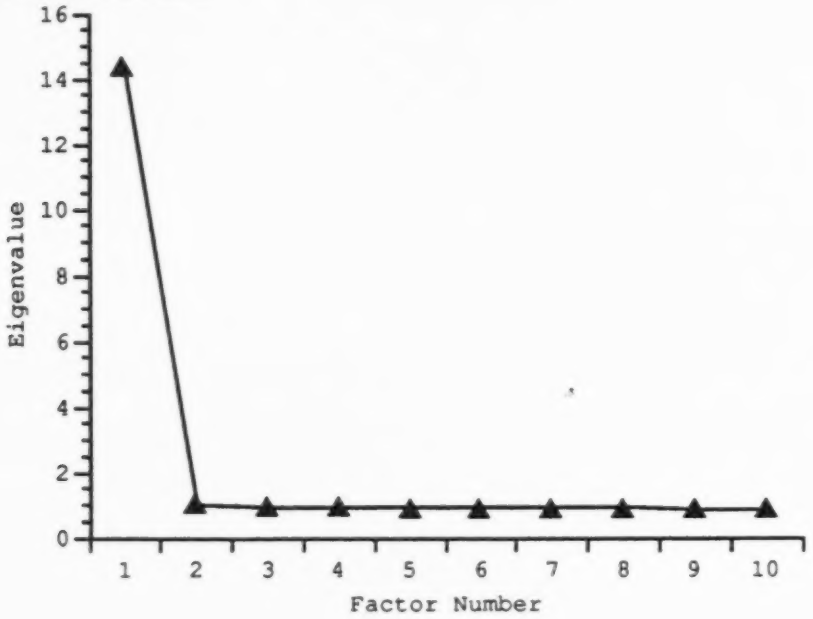
Factor	Lambda	Variance Proportion
1	14.392	.2875
2	1.106	.0204

Using Kaiser's criterion ($\lambda \geq 1$), two factors were identified and accounted for a total of 30.8 percent of the total test variance. Of the proportionate variance contributions, 93.3 percent is accounted for by the first factor with the remaining 6.7 percent being accounted for by the other significant factor. While the total variance accounted for is not high, it is not unlike that of other achievement tests (Hsu & Shermis, 1987). The scree plot from this factor analysis, displayed in Figure 1, suggests the predominance of the first factor. In so far as the first factor has the largest impact, the unidimensionality assumption was thought to have been met[2].

Another assumption for IRT calibration is related to the probable stability of the IRT estimates. The sample consisted of 726 Pakeha students, but only 63 Maori students. Typically item bias studies in the United States consist of samples in the thousands of students (for both majority and minority populations) and the estimates tend to be very stable.

[2]Some analysts might take issue with our assumption because of the presence of the other significant factor or because of the low total variance explained. It should be noted that the second significant factor only accounds for about two percent of the total test variance. A more rigorous procedure for testing unidimensionality has been described by Cook, Dorans, and Eignor (1988). This technique employs a confirmatory factor analysis of the second-order factors using LISREL, but was not applied to this data set.

Figure 1. Scree plot from the factor analysis of the Progressive Achievement Test (Mathematics--Form A, Part 6)



There is a general controversy surrounding minimum sample sizes sufficient for IRT parameter estimation. Lord (1980) has suggested that IRT calibrations proceed only with minimum sample sizes of 1000 for each group. Since then simulation studies have indicated that the parameter estimates are robust with significantly smaller samples (Hambleton & Swaminathan, 1985). While confidence can probably be placed in the parameter estimates for the Pakeha group, the same cannot be said for the parameter estimates of the Maori. Because of this we decided to use the calibrated data as a way to classify patterns of response rather than using it as a basis for conducting statistical tests. Moreover, results from classification technique might be compared to results obtained via calculation of the Mantel-Haenszel chi-square test statistic (Holland & Thayer, 1989) or the IRT-z^2 statistic (Lord, 1980).

Results

The data were analyzed with LOGIST V (Wingersky, Barton, & Lord, 1982) using the three-parameter logistic model. Four items (103, 129, 145, 147) from the Maori group would not converge because they produced definite, non-positive matrices.

Problems of this nature most commonly occur in LOGIST when either the sample size is small or an insufficient number of students actually get to the question. These problematic items were eliminated from further analyses[3]. This reduced the effective number of items to 46.

Because parameter estimates for both groups were made separately, it was necessary to equate the values of one group to the metric of the other. This was accomplished using a robust mean and sigma transformation technique described by Stocking and Lord (1983)[4]. Using this approach, a linear transformation of the form

$$B(\text{New, Transformed}) = R * B(\text{New}) + M,$$

$$A(\text{New, Transformed}) = A(\text{New})/R \qquad [2]$$

is found which places the focal group item-parameters on the scale of the reference group. The R (slope) and M (intercept) of this linear transformation are chosen to minimize the average squared difference between number right true scores on the common items for an arbitrary sample of examinees from the reference group (Stocking, 1986). This method is similar to one proposed by Haebara (1980) and Divgi (1980). Table 2 shows the equated parameter estimates for the two groups.

As a way to compare the conclusions that might be drawn from the pattern analysis, two additional tests were generated. First, a Mantel-Haenszel (MH) chi-square test statistic was calculated based on the hypothesis of no differences between proportions of Pakeha's and Maori's, across different ability levels, of obtaining a correct response. Using a format of

		Score on Studied Item		
		1	0	Total
Group	R	A_j	B_j	n_{Rj}
	F	C_j	D_j	n_{Fj}
	Total	m_{1j}	m_{0j}	T_j

[3]Actually the a,b, & c parameter estimates for the three items were calculated using a procedure described by Wingersky, Barton, & Lord (1982). However the procedure does not permit calculation of a standard error for the b parameter. The equating technique requires this value in order to make its biweighting calculations. As a result, the items were eliminated from further analysis.

[4]The program used to equate the two groups was written in SAS Interactive Matrix Language by Pam Diamond, The University of Texas at Austin. It is based on Stocking's TBLT (Transform B's Using a Least squares Technique) program (Stocking, 1986).

where

Table 2. Equated parameter estimates for Pakeha & Maori on PAT Mathematics (46 of 50 items).

	Pakeha			Maori		
Item #	A	B	C	A	B	C
101	0.6783	-1.3114	.1374	1.6575	-0.4313	.2574
102	0.5583	-1.1832	.1374	1.6575	-0.0582	.1287
104	0.4071	0.6729	.1374	0.2271	0.6712	.0834
105	1.3544	-0.1988	.2145	1.6575	0.2909	.1881
106	0.3197	-0.9307	.1374	0.2431	-0.4923	.0834
107	0.4606	0.4948	.1374	0.4998	1.0465	.1380
108	1.1682	0.3752	.2722	1.6575	0.2902	.1354
109	0.6708	0.1192	.1374	1.5860	0.2149	.0647
110	0.6061	0.4437	.0989	0.5173	0.1938	.0834
111	0.9256	0.8427	.1252	0.6910	0.1575	.1329
112	0.5756	0.2729	.1374	1.2481	0.3498	.0793
113	0.8166	-0.2877	.1374	1.6575	0.9568	.3678
114	0.8383	0.5030	.0764	1.6575	0.7321	.1043
115	1.1129	0.3938	.2028	1.6575	0.6904	.1942
116	0.5530	0.1343	.1374	1.6575	1.1481	.2381
117	1.5668	0.0333	.3061	1.3482	-0.0657	.0000
118	0.9008	0.2046	.1140	0.2750	1.1548	.0834
119	1.0513	0.4724	.1605	1.6575	0.7272	.2903
120	1.7500	0.8256	.0736	1.6575	1.2506	.1140
121	0.2893	0.6940	.1374	0.6274	0.7310	.0901
122	0.9760	0.6049	.1238	1.6575	0.9247	.1481
123	0.9458	0.6303	.0690	1.6575	1.3133	.2122
124	0.5914	-0.7843	.1374	1.0293	-0.3047	.0000
125	1.1540	-0.4025	.0484	1.6442	0.2835	.0000
126	0.8291	-0.9619	.1374	0.6613	-1.1618	.0834
127	1.3816	0.6981	.1973	1.6575	1.1432	.2735
128	0.9385	1.1069	.2205	0.1592	0.9845	.0834
130	0.4598	1.0676	.1374	0.8383	1.8145	.2303
131	0.8256	0.8294	.1825	0.6683	1.1632	.2420
132	0.3900	0.8799	.0155	0.4094	1.0171	.0834
133	0.8793	0.1424	.1008	1.2363	0.3606	.1911
134	0.5925	-0.4885	.1374	0.7590	-0.4537	.0000
135	0.8326	1.0874	.0830	0.8383	1.2882	.0623
136	0.5476	0.8680	.0161	1.4028	1.0961	.1215
137	0.9362	0.9817	.1081	0.2093	3.5531	.0000
138	0.8930	-0.1567	.0000	1.6575	0.8481	.2679
139	1.3312	0.8722	.1220	1.6575	1.5292	.1034
140	0.9867	1.4898	.1954	1.6575	1.5262	.1708
141	1.1176	0.7791	.1703	0.5130	1.0498	.0834
142	0.7411	-1.7531	.1374	1.6525	0.1355	.4324
143	0.9741	0.8631	.1339	1.0869	0.6714	.0000
144	1.5310	1.9539	.2182	1.1336	1.5150	.1781
146	0.5947	0.7671	.1060	1.6575	1.7408	.2996
148	0.5271	-0.4513	.1374	0.2768	-0.1573	.0834
149	1.7500	1.1718	.2630	1.0940	1.5159	.1123
150	1.1964	1.2891	.2120	0.1154	3.5620	.0000

R stands for the reference group, F stands for the focal group, and j is the matching or stratification variable (e.g., matching on a test score).

The MH chi-square statistic can be expressed as

$$\frac{\sum_{1}^{j}\left(A_j - \sum_{1}^{j} E(A_j)\right)^2}{\sum_{1}^{j} Var(A_j)} \qquad [3]$$

where $E(A_j) = n_{Rj}m_{1j}/T_j$, and

$$\sum_{1}^{j} Var(A_j) = \frac{n_{Rj}n_{Fj}m_{1j}m_{0j}}{T_j^2(T_j - 1)}$$

Most computer programs which calculate the MH statistic apply a continuity correction to improve the accuracy of the chi-square percentage points as approximations to the observed significance levels as shown in Formula [4] (Holland & Thayer, 1989).

$$\frac{\left(\left|\sum_{1}^{j} A_j - \sum_{1}^{j} E(A_j)\right| - \frac{1}{2}\right)^2}{\sum_{1}^{j} Var(A_j)} \qquad [4]$$

Because of the relatively small size of the focal group, total test scores were divided into three strata on which both the reference and focal groups were matched: (high) 67-99+ percentile, (medium) 34-67 percentile, and (low) 1-33 percentile. Scheuneman (1979) found acceptable results by matching on a criterion variable using 3-5 intervals. While it would have been more desirable to match on each test score (range:4-49), the number of required intervals would have precluded a reliable chi-square statistic. The MH test statistics were calculated using SYSTAT 5.0™.

In addition to calculating the MH statistics, the *z* test discussed by Lord (1980) was also calculated on the IRT b parameter. This test takes the form

$$z = \frac{(b_r - b_f)}{SE(b_r - b_f)} \qquad [5]$$

where

$$SE(b_r - b_f) = \sqrt{var\ b_r + var\ b_f}$$

Table 3. Chi-square values for the Mantel-Haenszel and z^2 test examining bias on the Progressive Achievement Tests (Mathematics--46 of 50 items).

Item #	MH χ^2	z^2
101	0.00	5.14
102	3.19	7.79**
104	0.20	0.00
105	1.39	4.47*
106	0.86	0.03
107	1.01	0.55
108	0.03	0.12
109	0.09	0.14
110	1.89	0.18
111	0.16	1.86
112	0.23	0.06
113	2.22	13.90***
114	0.28	1.20
115	0.58	1.64
116	5.93*	6.57**
117	0.61	0.35
118	0.40	0.58
119	1.07	0.95
120	0.20	2.21
121	1.47	0.00
122	0.75	1.90
123	0.20	3.80
124	0.80	1.80
125	7.25**	21.05***
126	0.54	0.09
127	0.00	2.03
128	3.40	0.00
130	1.01	0.58
131	0.50	0.28
132	0.35	0.03
133	0.98	0.55
134	0.02	0.01
135	0.27	0.20
136	0.39	0.57
137	2.31	0.51
138	0.27	16.63***
139	4.23*	3.02
140	0.78	0.01
141	0.17	0.21
142	3.56	19.10***
143	0.06	0.77
144	0.13	0.68
146	0.12	2.76
148	0.04	0.03
149	4.01*	0.48
150	0.11	0.24

*p <.05
**p <.01
***p <.001

Oftentimes, z^2 is referred to a table of the chi-square distribution with 1 degree of freedom as a way to maintain comparability with other tests based on other chi-square tests based on a single-degree-of-freedom. Both the results from the Mantel-Hanszel chi-square test and the z^2 test are given in Table 3.

The Mantel-Haenzel chi-square test identified four items which differed significantly: 116, 125, 139, and 149. The z^2 test identified seven item which were significantly different: 102, 105, 113, 116, 125, 138, 142. In examining the items which differed significantly across both tests (116, 125), the chi-square value for the z^2 test was much higher than for the MH test. This may have been due in part to the fact that the groups were matched on a rather gross level of test performance (high, medium, low) or that the standard error of b's for the focal group were high.

Next, based on the equated item parameters, ICC curves were generated for both groups. The curves are based on Formula [1] and represent a continuum of θ. However actual computations were based on discrete values of θ ranging from -3 to +3. The 46 ICCs were classified into 6 patterns. Figures 2-7 show the item numbers and associated ICC patterns. In parentheses after each item is the classification for the item type (e.g., Numeration).

As Figures 2-3 demonstrate, the general trend was that for a given level of ability the probability of getting an item correct was greater for the Pakeha group than for the Maori group. This trend was clear for 15 of the 46 items (33 %). Figure 2, for example, shows items in which the probability of getting the item correct is higher for the Pakeha group across the entire ability spectrum. Figure 3 is very similar to Figure 2 except that items from this pattern showed that at one or both ends of the ability spectrum the probability of getting the item correct was approximately the same for the two groups. The ICCs usually converged towards the upper end of the ability distribution.

There were exceptions to this trend, but in most cases the exception involved only a limited portion of the ability distribution. For example, with the items exemplified by Figure 4, Maoris in the lower end of the ability distribution had a higher probability of getting an item correct than Pakeha students. In the middle

and upper portion of the ability distribution Maoris had a lower probability of getting the item right. Conversely, with the items like those shown in Figure 5, Maoris in the upper end of the ability distribution had a higher probability of getting an item correct than did Pakeha students. In the middle and lower portion of the ability distribution Maoris had a lower probability of getting the item correct.

Figure 6 reflects a curious distribution of ICCs for the Pakeha and Maori groups. For three of the four items, Pakeha students in the middle of the ability distribution had a higher probability of getting the item correct than did Maori students. Maoris in the upper and lower ends of the ability distribution had a lower probability of getting the item right. On one of the four items, the reverse pattern was observed.

Item 133 was the only item in which the ICCs were essentially the same for both groups. Figure 7 shows that the Pakehas have a higher probability of getting the item correct across the ability distribution, though the actual differences between the curves are very small.

Discussion and Conclusions

Because the data were drawn from a real sample and not a simulated set of test scores, it is impossible to know what the true parameters actually are. In addition, the interpretations of the z^2 test results and ICC curves may be hampered by an insufficient sample size of the focal (Maori) group. However, the following trends seem to emerge: the pattern classification analysis, based on the ICC curves, would lead one to conclude more bias than would either the results from the z^2 or MH tests. The MH procedure, not based on IRT parameters, seems to be the most conservative approach to evaluating test bias, though there was 85 percent agreement between the conclusions one would arrive at using the MH procedure and those derived from the results of the z^2 test.

Figure 2. Items typified by the following ICCs (Taken from Item 143)

Item	Item
107 (Geometry)	117 (Logic)
130 (Numeration)	135 (Sets)
139 (Number Operations)	140 (Sets)
143 (Measurement)	149 (Number Operations)

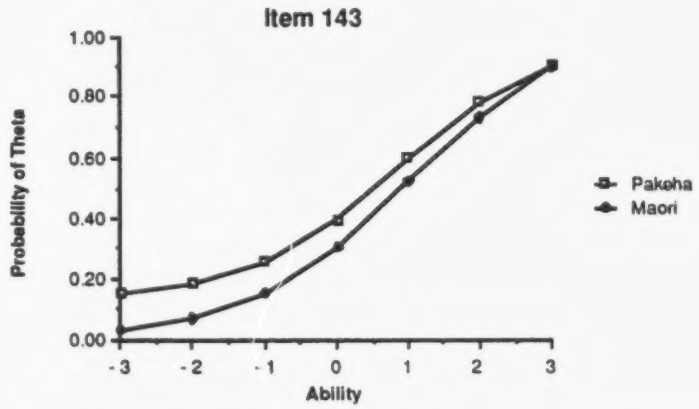


Figure 3. Items typified by the following ICCs (Taken from Item 134)

Item	Item
105 (Sentences)	106 (Measurement)
108 (Number)	115 (Statistics)
122 (Measurement)	125 (Logic)
134 (Number Operations)	

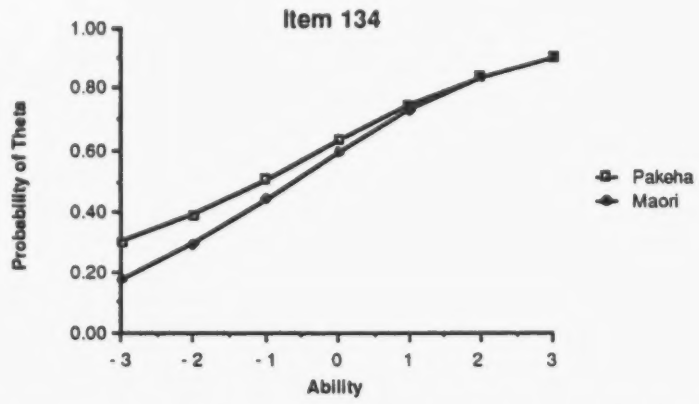


Figure 4. Items typified by the following ICCs (Taken from Item 120)

Item		Item	
104	(Number Operations)	110	(Number Operations)
111	(Number Operations)	113	(Number Operations)
118	(Number)	120	(Number Operations)
123	(Geometry)	126	(Sets)
127	(Numerations)	128	(Number Operations)
131	(Number Operations)	132	(Statistics)
137	(Geometry)	138	(Measurement)
141	(Sentences)	146	(Number)
148	(Geometry)	150	(Spatial Relations)

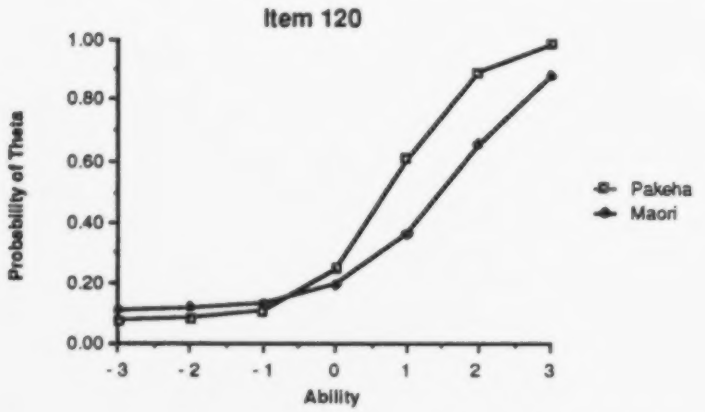


Figure 5. Items typified by the following ICCs (Taken from Item 109)

Item		Item	
101	(Measurement)	102	(Number Operations)
109	(Numeration)	112	(Sentences)
114	(Number Operations)	116	(Geometry)
121	(Sentences)	124	(Spatial Relations)

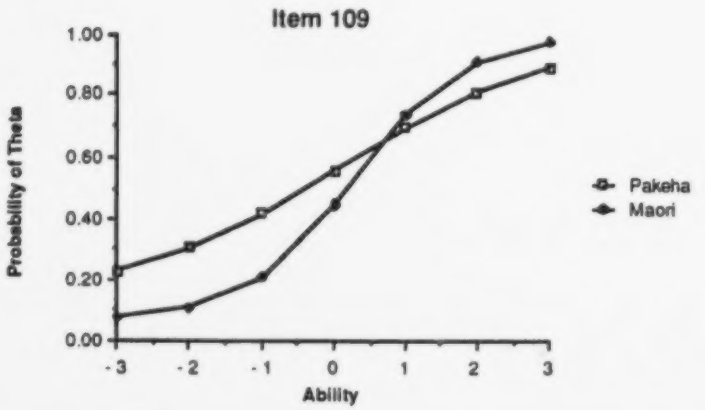


Figure 6. Items typified by the following ICCs (Taken from Item 142)

Item	Item
119 (Sets)	136.(Number)
142 (Spatial Relations)	144.(Geometry)

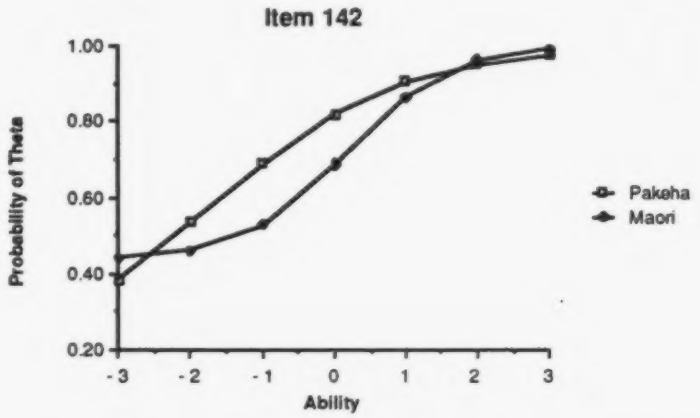
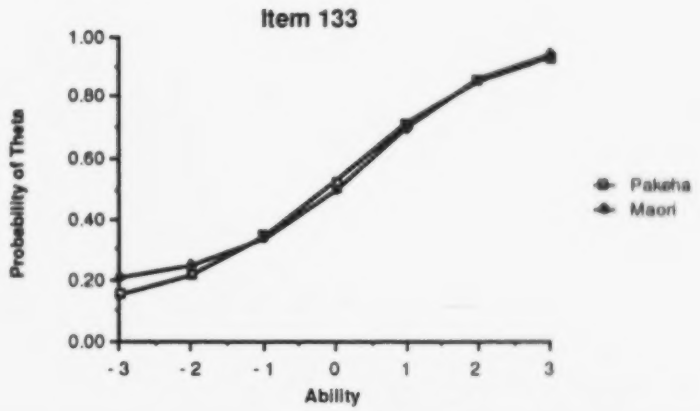


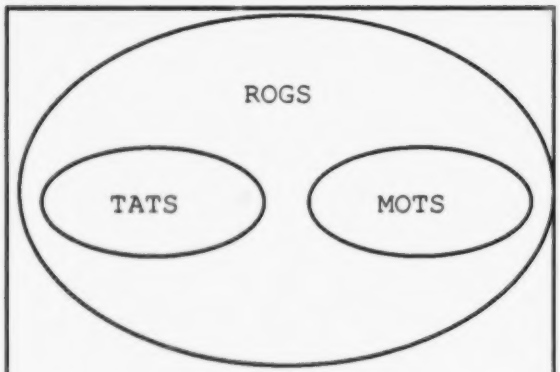
Figure 7. Items typified by the following ICCs (Taken from Item 133)

Item
133 (Spatial Relations)



All three approaches identified items 116 and 125 as being potentially biased. It should be noted that the ICC curves of these two items were placed into separate pattern classifications. There are a number of plausible reasons for why bias might exist. For example, on item 125 (Logic), the ICC produced was similar to that of Figure 3 above (see Appendix A for the graph of the actual item). The item reads:

Figure 8. The text of item 125



This diagram shows that

(A) all mots are rogs.
(B) all rogs are mots.
(C) all mots are tats.
(D) no rogs are tats.
(E) no mots are rogs.

The pattern classification suggests that the Pakeha low-ability level group outperforms the Maori low-ability level group, but that both Pakehas and Maoris do equally well at the upper-ability levels. Performance on this type of item may be reflective of schooling effects, the specific nature of which will be explored in future studies.

While pattern classification analysis based on ICC curves is problematic in situations where the size of the focal group is limited, it may nevertheless be useful as a supplemental technique to explore the nature and cause of group differences identified by non-IRT methods. Moreover, there appears to be high agreement between conclusions based on both non-IRT and IRT methods. IRT methods in this situation may have, in part, been more sensitive to group differences because of the relatively large standard errors of the focal group resulting from the small sample size used in calibrating the item parameters.

References

Beck, L.R., & St. George, R. (1983). The alleged cultural bias of PAT Reading Comprehension and Reading Vocabulary Tests. *New Zealand Journal of Educational Studies, 18*, 32-47.

Chapman, J.W., St. George, R, & Ibell, R. (1985). Error rte differences in teacher marking of the Progressive Achievement Tests: Sex, ability, SES and ethnicity effects. *New Zealand Journal of Educational studies, 20*, 165-169.

Cook, L.L, Dorans, N.J., & Eignor, D.R. (1988). An assessment of the dimensionality of three SAT-verbal test editions. *Journal of Educational Statistics*, 13(1), 19-43.

Divgi, D.R. (1980, April). *Evaluation of scales for multilevel test batteries.* Paper presented at the Annual Meetings of the American Educational Research Association, Boston, MA.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3).

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Holland, P.W., & Thayer, D.T. (1989). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hsu, T.C., & Shermis, M.D. (April,1987). *Assessing psychometric qualities of a computer adaptive placement testing system.* Paper presented at the annual meetings of the American Educational Research Association, Washington, DC.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: Erlbaum.

Mellenbergh, G.J. (1988). *Experimental research on the explanation of item bias.* Paper presented at the annual meetings of the American Educational Research Association, New Orleans, LA.

Muthen, B.O., Kao, C, & Burstein, L. (1988). *Instructional sensitivity in mathematics achievement test items: Application of a new IRT-based detection technique*. Paper presented at the annual meetings of the American Educational Research Association, New Orleans, LA.

Reid, N.A., & Hughes, D.C. (1974). *Progressive Achievement Tests: Teacher's manual (Mathematics)*. Wellington: New Zealand Council for Educational Research.

Scheuneman, J.D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24*, 97-118.

St. George, R. (1987). Applying an internal criterion of test bias to Progressive Achievement Tests and the Test of Scholastic Abilities: Item x group interactions. *New Zealand Journal of Educational Studies, 22*, 117-119.

St. George, R., & Chapman, J.W. (1987). *Item analysis of Progressive Achievement Tests at the Form 2 class level*. Wellington, New Zealand: Department of Education (Contract Report # 41/17/149).

Stocking, M. (1986). *Documentation for transforming b's using a least squares technique*. Princeton, NJ: Educational Testing Service.

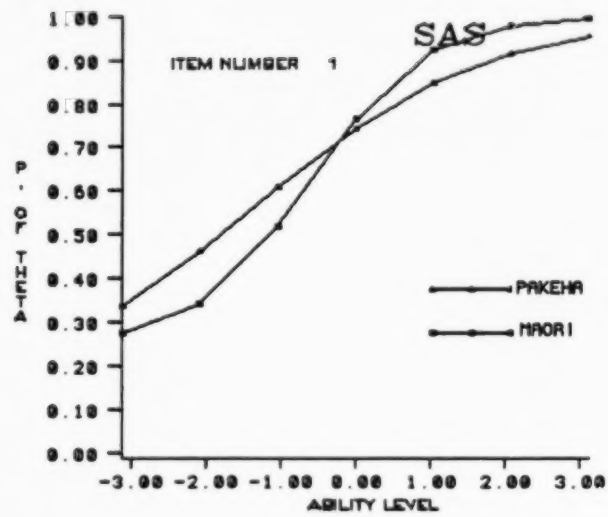
Stocking, M., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

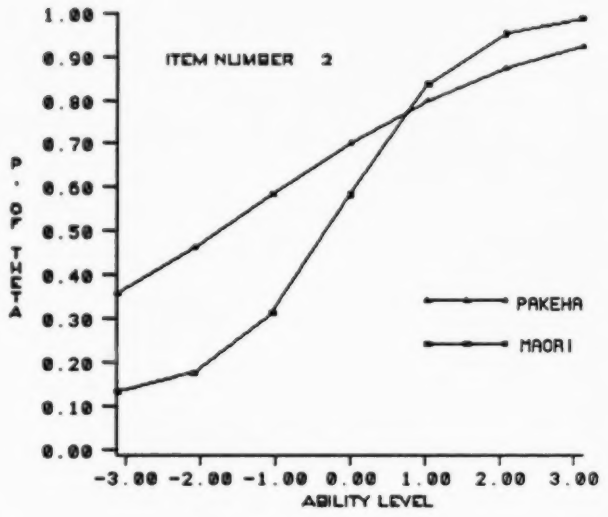
Thissen, D., Steinberg, L., & Wainer, H. (1989). Use of item response theory in the study of group differences in trace lines. In H. Wainer and H.I. Braun (Eds), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

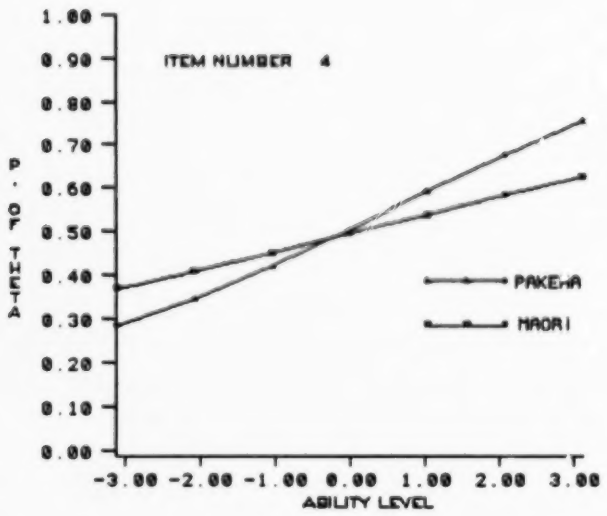
Wilkinson, L. (1989). *SYSTAT: The System for Statistics*. Evanston, IL: SYSTAT, Inc.

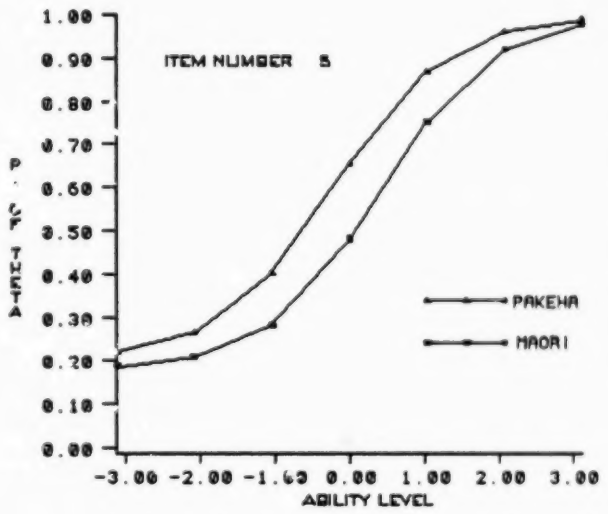
Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). *LOGIST user's guide*. Princeton, N.J.: Educational Testing Service.

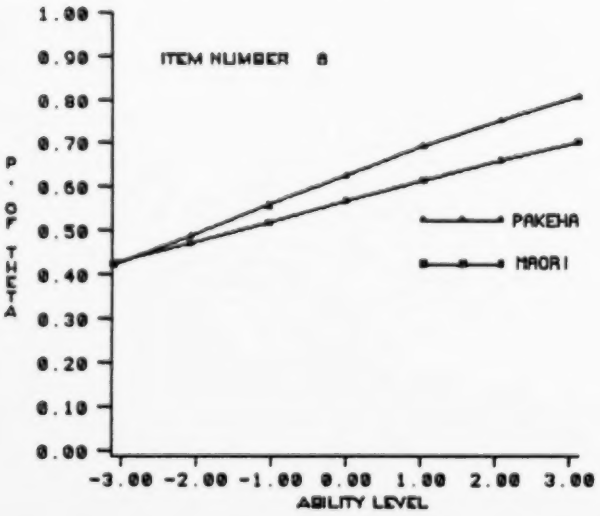
Appendix A

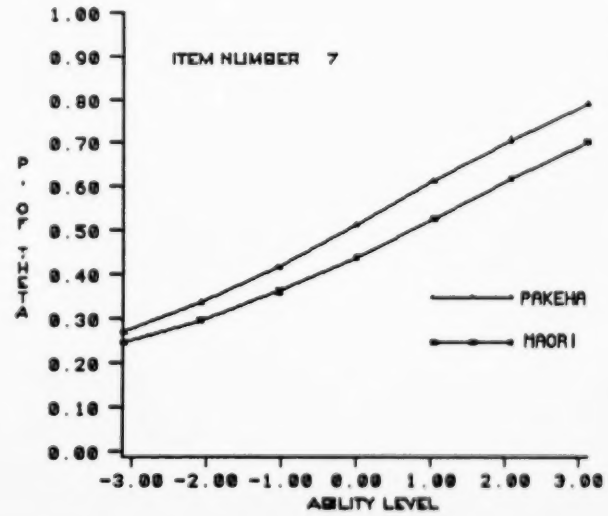

SAS
ITEM NUMBER 1
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

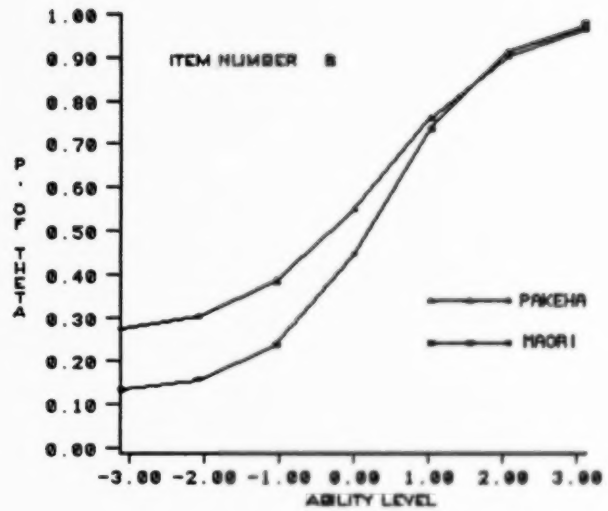

ITEM NUMBER 2
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

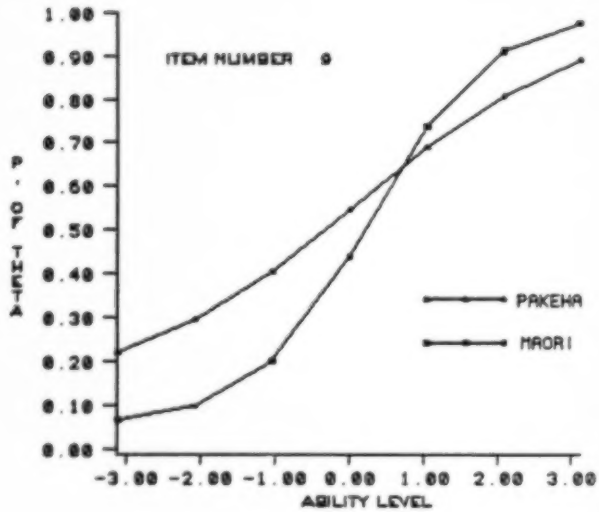

ITEM NUMBER 4
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

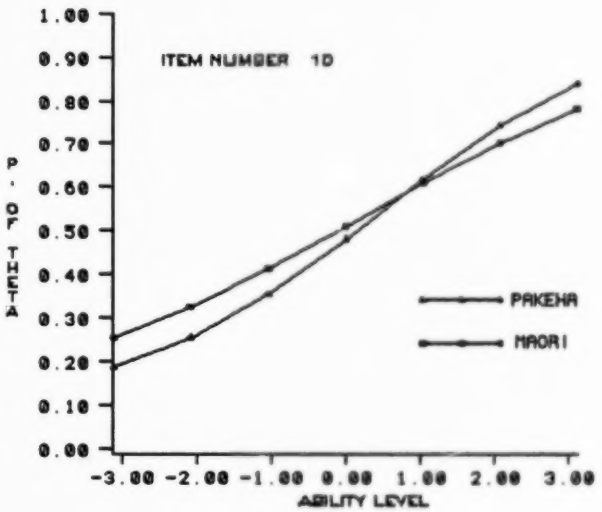

ITEM NUMBER 5
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

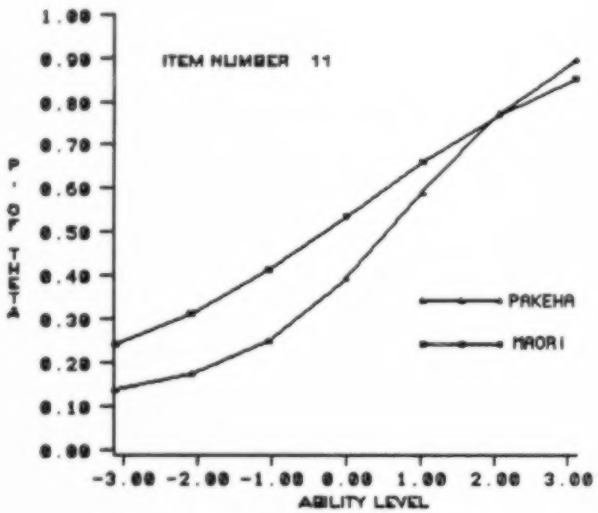

ITEM NUMBER 6
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

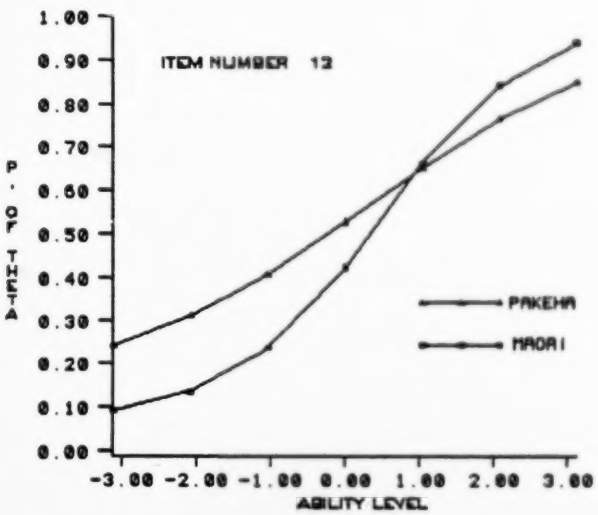

ITEM NUMBER 7
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

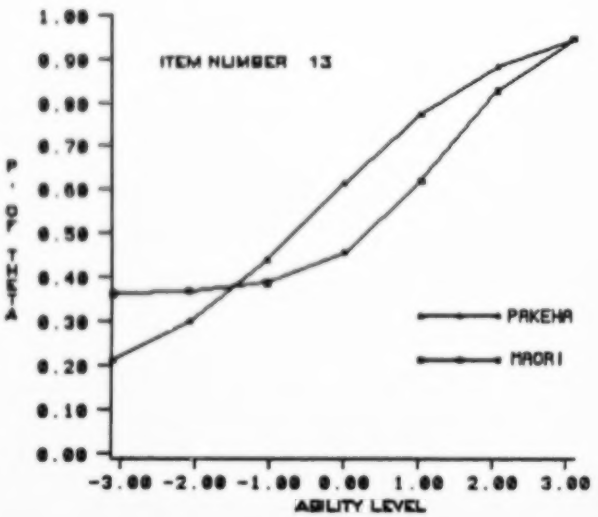

ITEM NUMBER 8
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

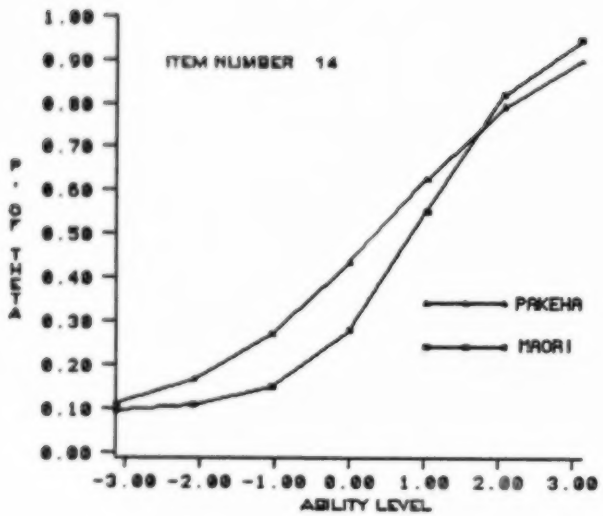

ITEM NUMBER 9
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

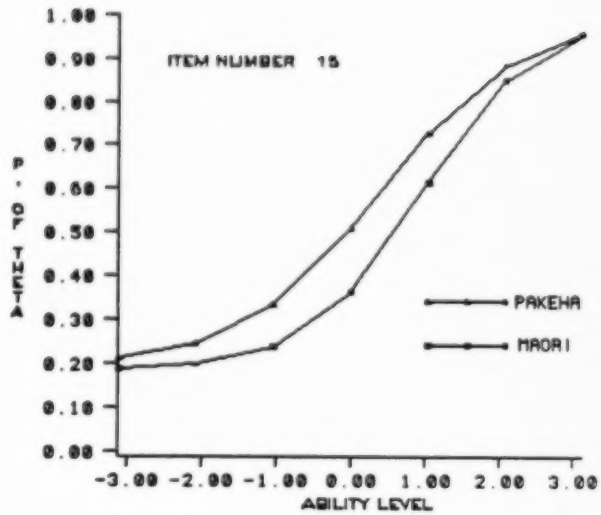

ITEM NUMBER 10
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

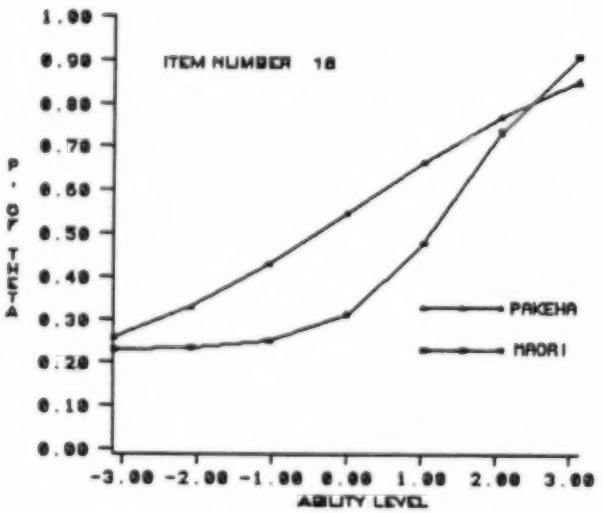

ITEM NUMBER 11
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

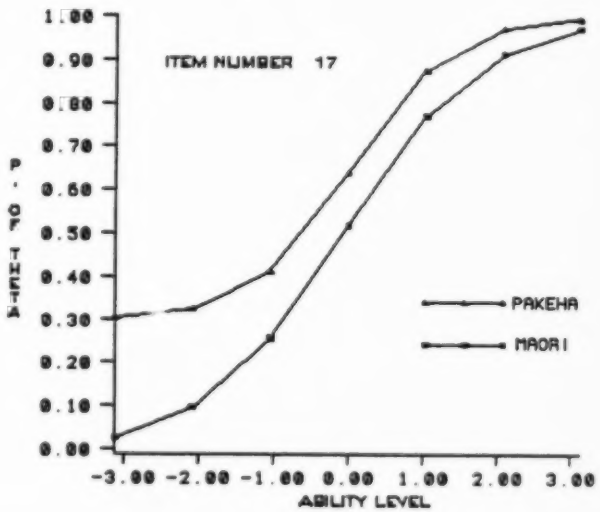

ITEM NUMBER 12
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

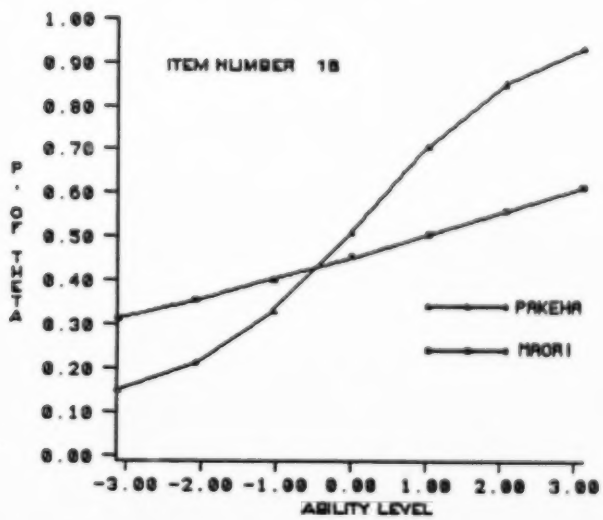

ITEM NUMBER 13
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

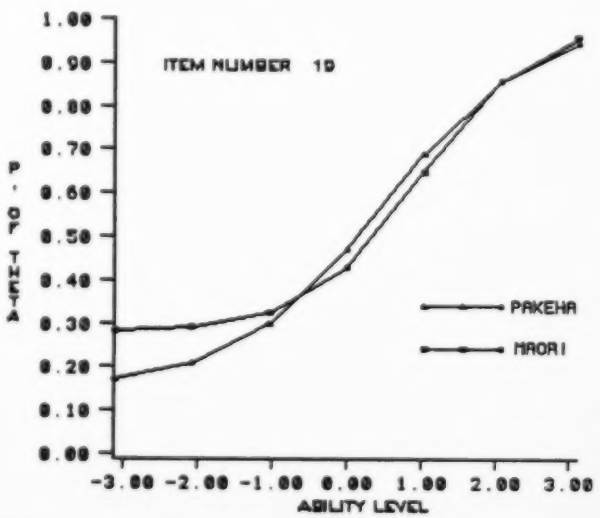

ITEM NUMBER 14
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

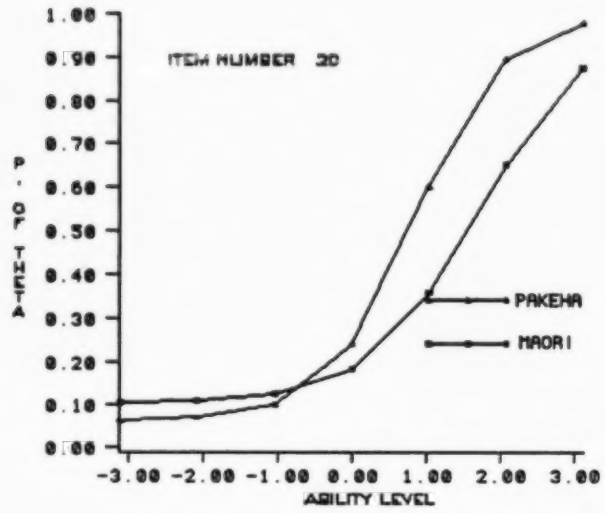

ITEM NUMBER 15
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

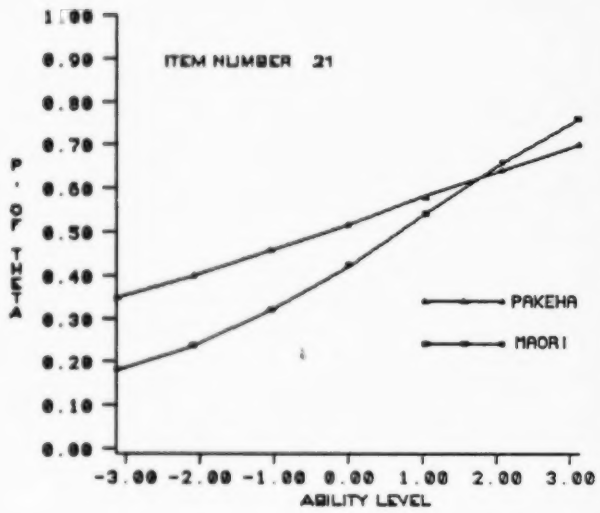

ITEM NUMBER 16
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

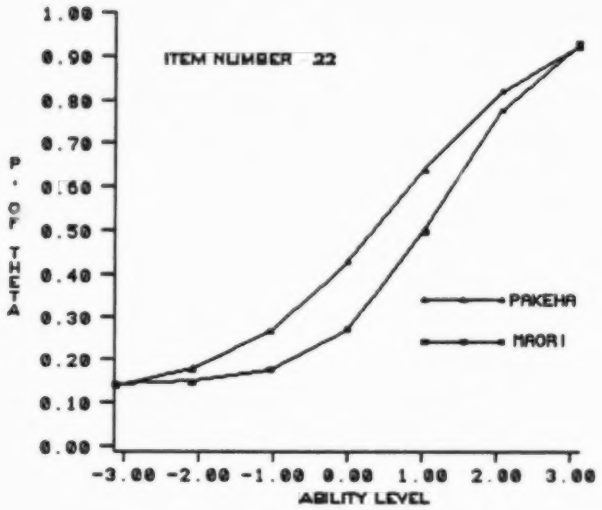

ITEM NUMBER 17
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

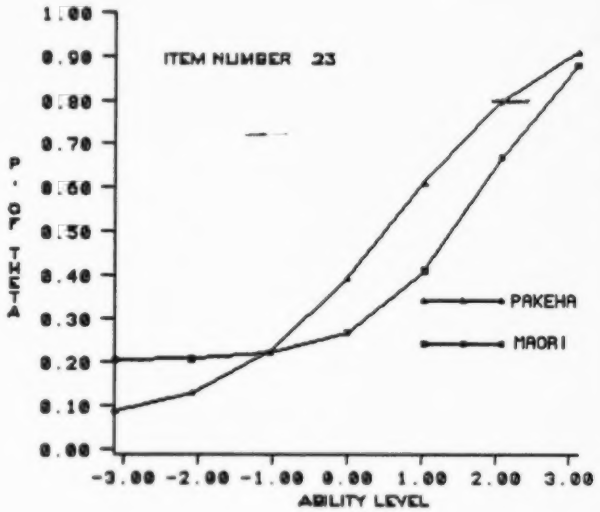

ITEM NUMBER 18
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

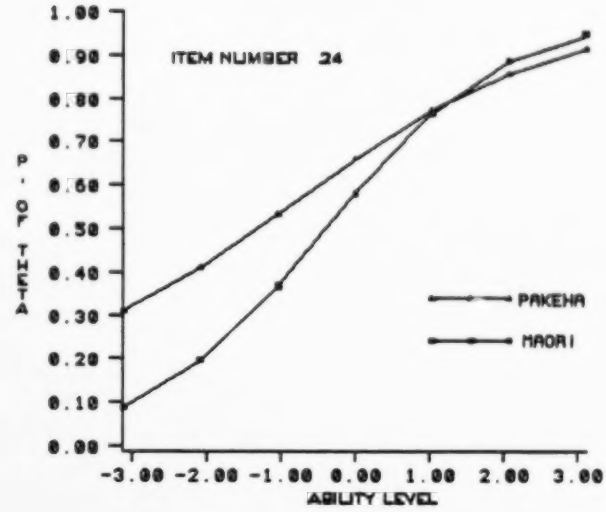

ITEM NUMBER 19
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

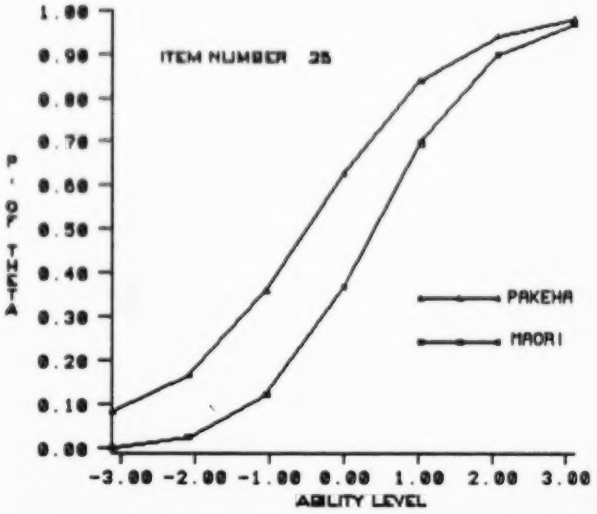

ITEM NUMBER 20
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

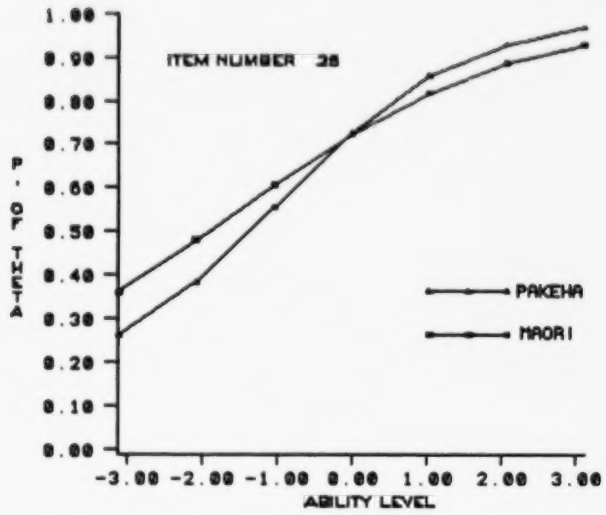

ITEM NUMBER 21
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

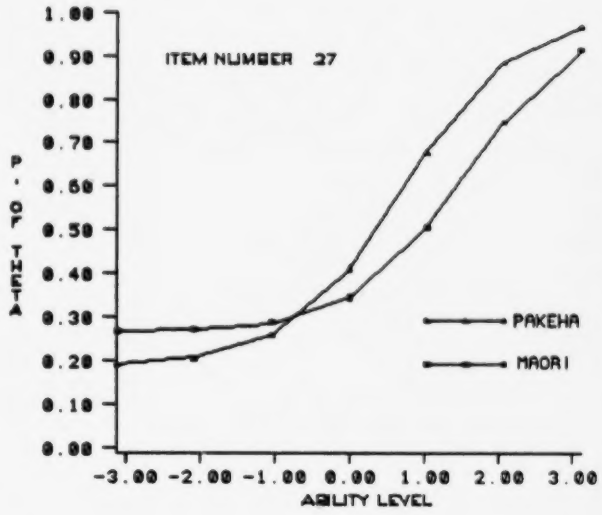

ITEM NUMBER 22
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

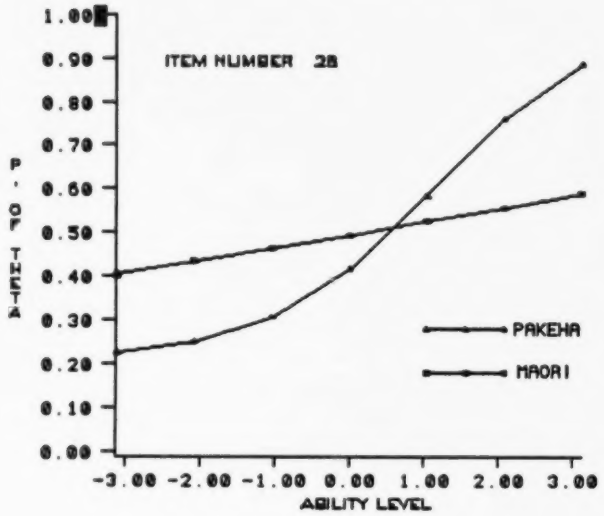

ITEM NUMBER 23
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

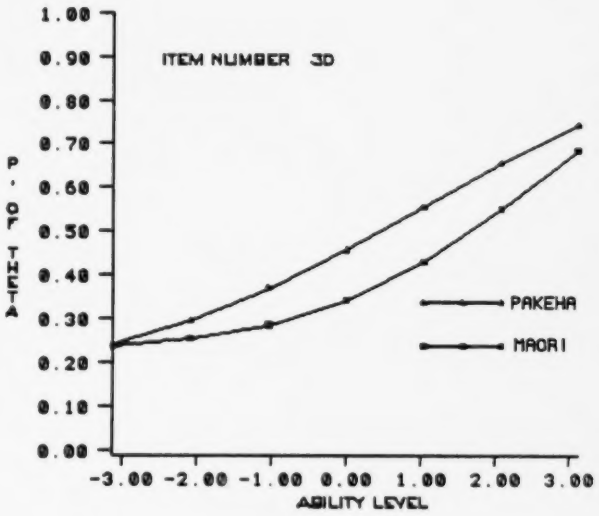

ITEM NUMBER 24
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

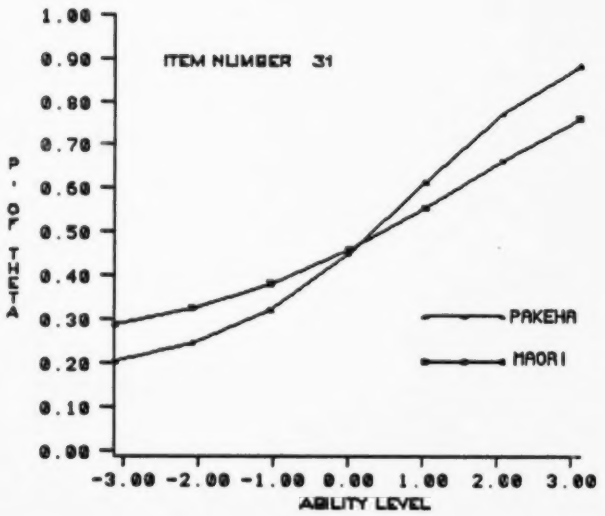

ITEM NUMBER 25
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

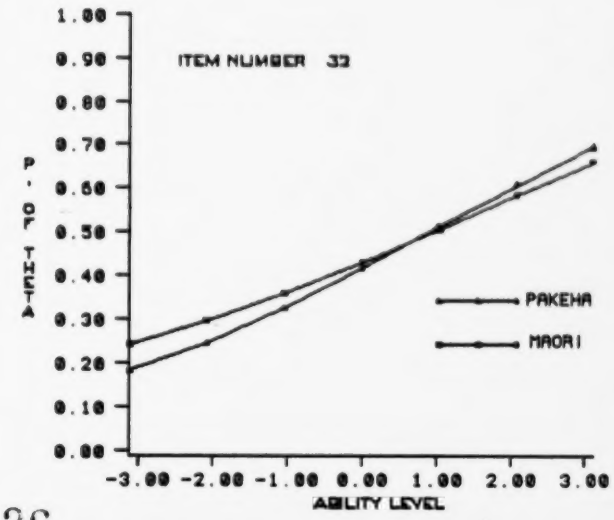

ITEM NUMBER 26
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

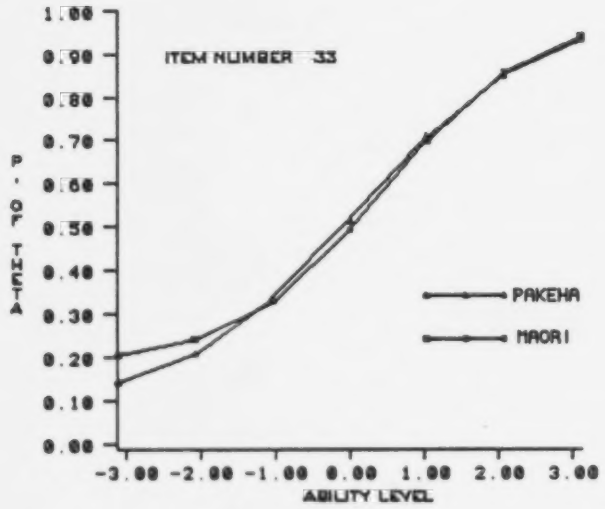

ITEM NUMBER 27
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

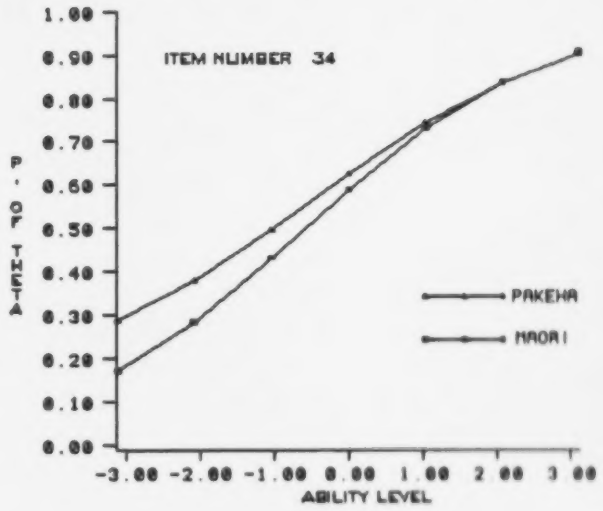

ITEM NUMBER 28
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

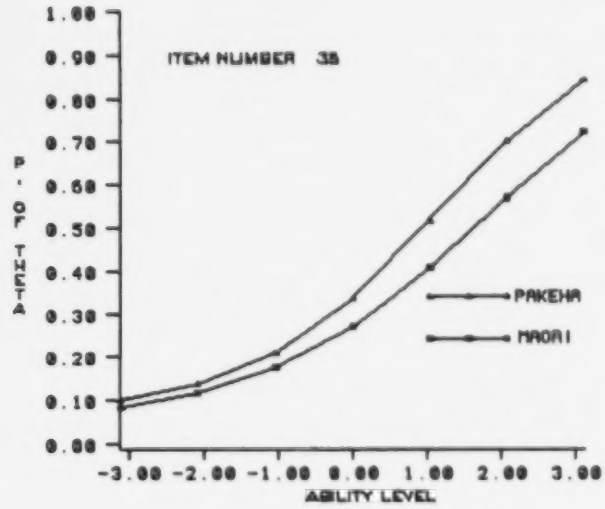

ITEM NUMBER 30
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI


ITEM NUMBER 31
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

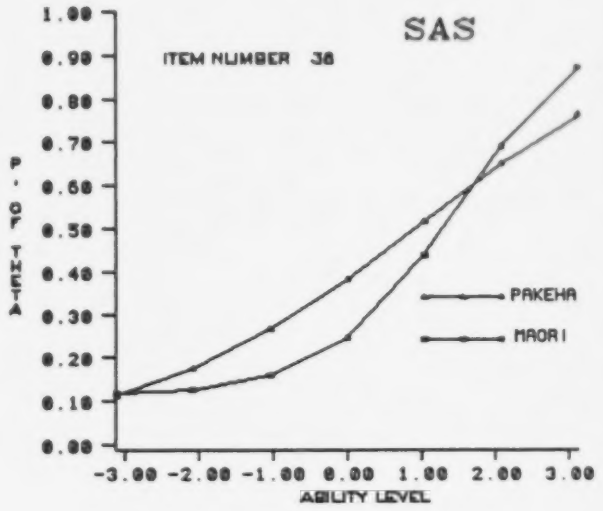

ITEM NUMBER 32
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

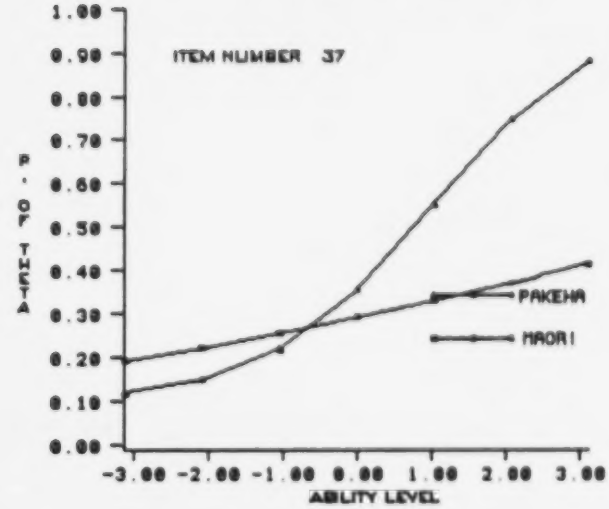

ITEM NUMBER 33
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

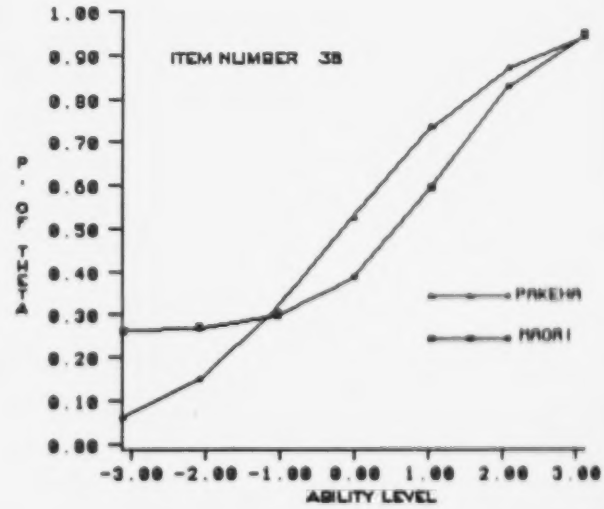

ITEM NUMBER 34
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

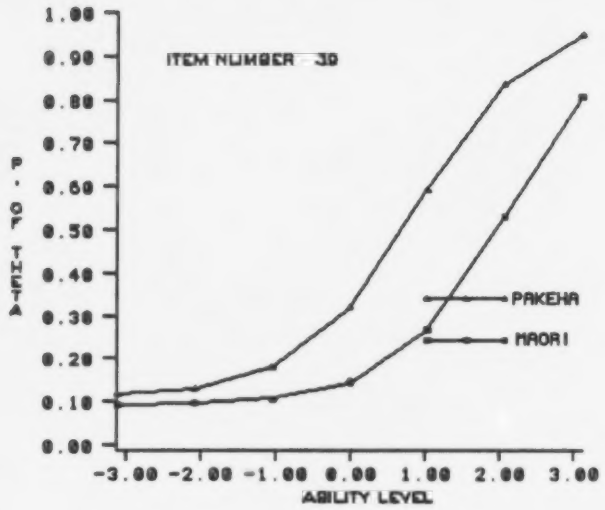

ITEM NUMBER 35
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

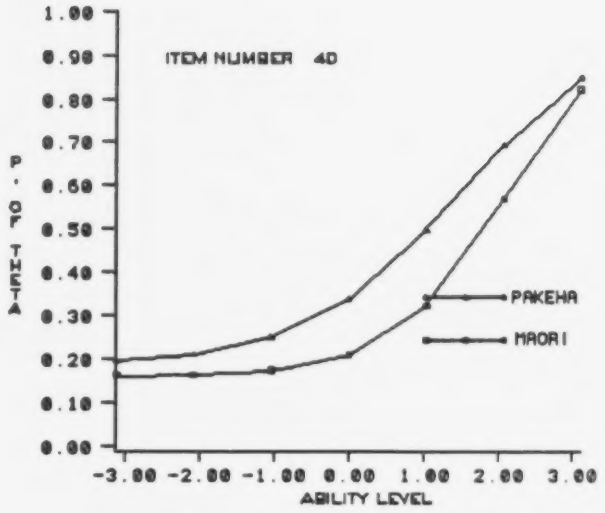
SAS

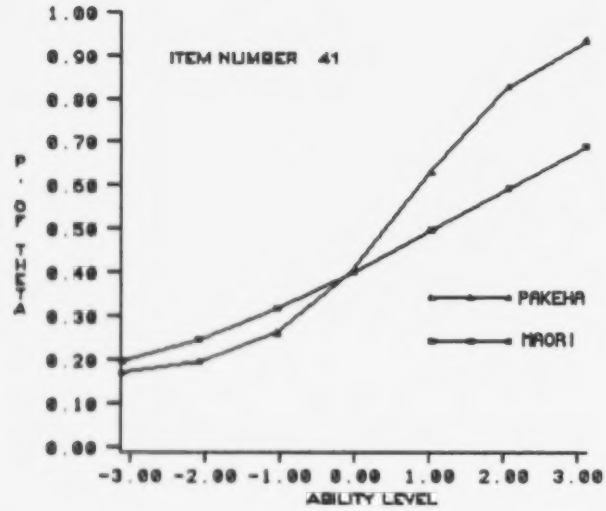

ITEM NUMBER 36
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

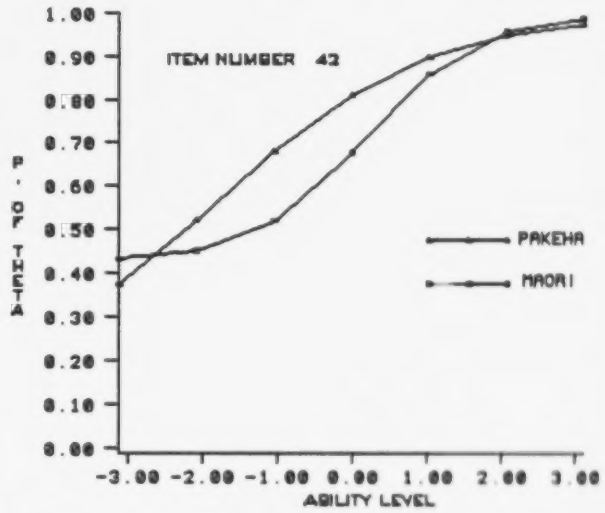

ITEM NUMBER 37
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

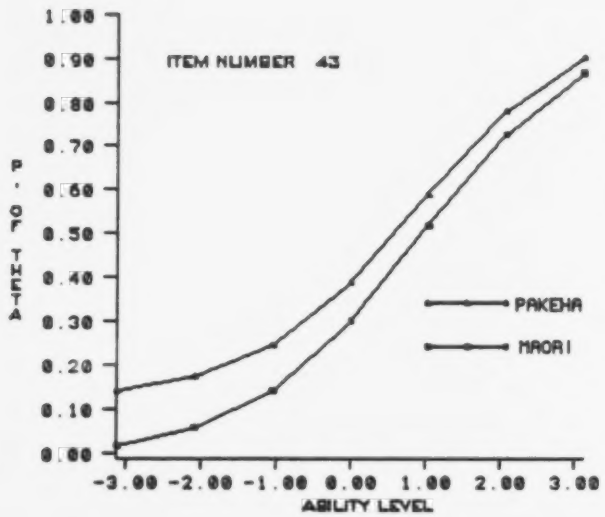

ITEM NUMBER 38
P - OF THETA
ABILITY LEVEL
PAKEHA
MAORI

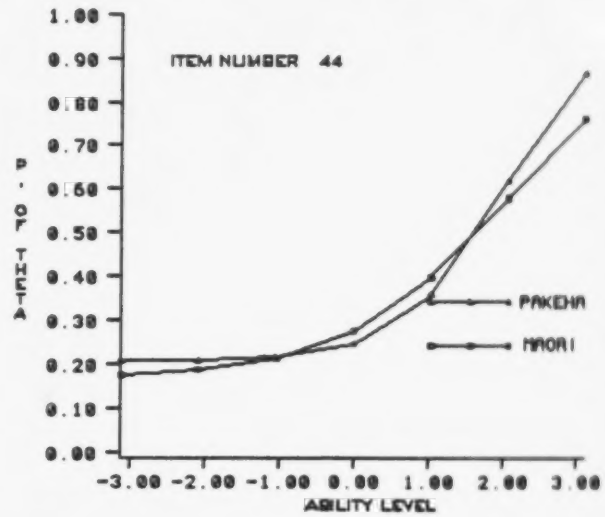

ITEM NUMBER 39
P · OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

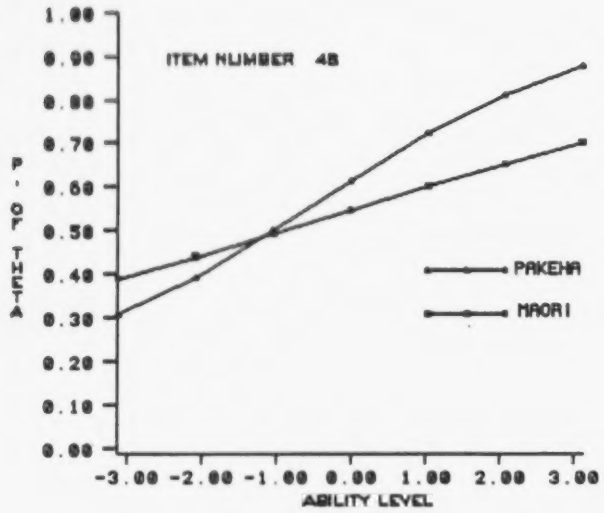

ITEM NUMBER 40
P · OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

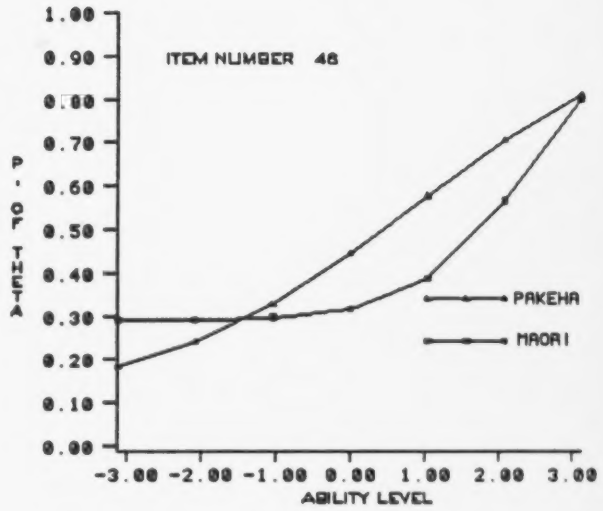

ITEM NUMBER 41
P · OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

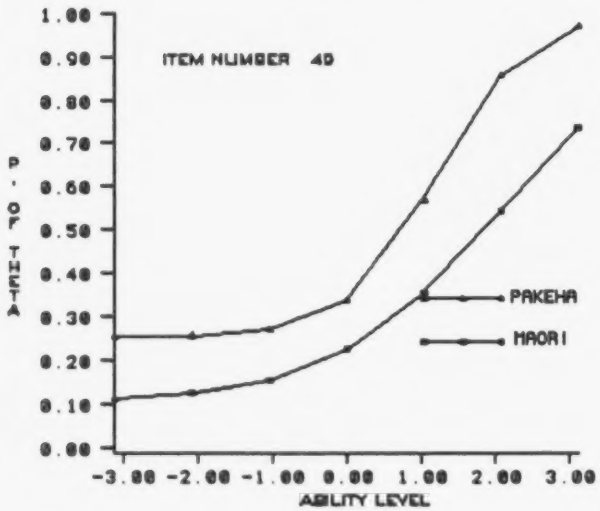

ITEM NUMBER 42
P · OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI

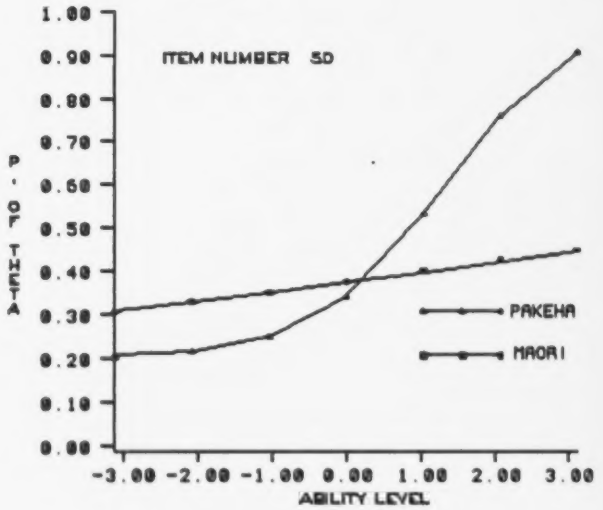

ITEM NUMBER 43
P · OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI


ITEM NUMBER 44
P · OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI


ITEM NUMBER 48
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI


ITEM NUMBER 46
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI


ITEM NUMBER 49
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI


ITEM NUMBER 50
P - OF THETA
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ABILITY LEVEL
PAKEHA
MAORI